DOCUMENT RESUME

ED 171 104 FL 010 034

AUTHOR Dixon, J. E. G.
TITLE Concordances KWIC and Complete: An Appraisal.
PUB DATE 78
NOTE 7p.
AVAILABLE FROM Association for Literary and Linguistic Computing, Sidgwick Site, Cambridge CB3 9DA, England (£4 or US$8 for entire Bulletin)
JOURNAL CIT Association for Literary and Linguistic Computing Bulletin; v6 n1 p28-33 1978

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Automatic Indexing; Citation Indexes; *Computational Linguistics; *Concordances; Coordinate Indexes; *Indexes (Locaters); Information Retrieval; Literary Analysis; Literature; *Permuted Indexes; Reference Materials; *Research Tools; Structural Analysis; Word Frequency
IDENTIFIERS *KWIC Index

ABSTRACT

The application of computer techniques to the making of concordances and indexes has brought about many innovations including the KWIC format and the exhaustive concordance. Although claims for the merits of these concordances are not always justified, the KWIC method offers the following advantages: (1) the detection of formulaic expressions and repeated word patterns; (2) simplicity; (3) speed; (4) maximization of computer use; and (5) minimization of the editor's role. The traditional concordance was designed by philologists and is problematic not only because of its sheer bulk, but also because it tends to assume that word frequency is directly proportional to unimportance and it tends to favor only literary interests, almost neglecting sizeable linguistic interests. Solutions to size and economic problems are: (1) to make tapes available to researchers who have access to a computer; (2) to print references only in an appendix to the main concordance; (3) to issue sub-concordances on microfilm or microfiche; and (4) to publish as cheaply as possible a series of sub-concordances as fascicules, each containing pages according to specific categories and criteria. Traditionally a tool of philologists, concordances are perhaps not the best research instrument for linguists. But the concordance, improved by the refinements and analytical devices provided by the editor and computer together, will continue to serve the philologist well in the future. (Author/MHP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

CONCORDANCES KWIC AND COMPLETE: AN APPRAISAL

J.E.G. Dixon *(University of Winnipeg, Canada)*

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. L. Dawson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

* * * *

Abstract

This article examines the claims made for the merits of the KWIC concordance and the exhaustive concordance. The author draws the conclusion that the claims are not always justified, and that where they are, other ways of realizing the objectives sought may be preferable, since the disadvantages of these concordances outweigh the value of their distinctive features. The criteria of judgement applied to these, as to all concordances, are the principles of scholarship, and the essentially practical ones of usability, economics, and the user envisaged.

* *

The application of computer techniques to the making of concordances and indexes has brought about many innovations in these enterprises. Two features of computer-generated concordances which were impossible or impracticable in the days of the manual compiler are the KWIC format and the exhaustive concordance. Both have found their practitioners and apologists. In this article I wish to examine the pros and cons of these two features.

KWIC - or what?

In the KWIC type of concordance the entry words are arranged in a vertical column down the centre of the page. The citations extend to either side of them in equal measure, and reproduce as much text as the page size will allow.[1] Further, the words or word groups to the right of the keyword, for all occurrences of each keyword, are sorted in alphabetical order.[2] The reasons and justifications given for the arrangement chosen are:

> ... it has the great advantage of grouping identical phrases together. Besides making it a simple matter to locate a particular collation of words, it brings to light parallel phraseology which might go unnoted if the related entries were distributed over several pages (Packard (1968), Introduction, p. v).

> This method was adopted to enable users to ascertain the position of each entry with ease and to facilitate contextual analysis (cf. *to* - infinitive and *for to* - infinitive)
> (Kato (1974), Preface, p. vii).

> Ce procédé a été choisi pour plusieurs raisons: d'abord parce qu'il permet de faire une économie de place ...; ensuite parce que, s'agissant d'un texte en prose, nous étions forcés de délimiter nous-mêmes le contexte à envisager: le systeme KWIC permet d'équilibrer les contextes à gauche et à droite du mot-vedette. Nous donnons tous les mots du texte original, classés d'abord par ordre alphabétique, puis par ordre alphabétique des mots du contexte de droite .... Cette façon de faire

ED171104

FL010034

nous permet de retrouver bon nombre de constructions figées, syntaxiques ou stylistiques, tours spécifiques de l'auteur, formules, etc. (Hanon (1974), Préface, pp. 9-10).[3]

We may question the first two reasons given by Hanon, above. In the first place, if 'une économie de place' is achieved, it is because there are no spaces between lines of print, not because of KWIC; secondly, every editor is required to delimit the citations, and the manner in which it is done in this concordance negates the first reason (space) because of their inordinate length (see below). In Kato's work the part of the citation to the right of the keywords is neither alphabetized nor entered according to grammatical categories, so the user has to search for the infinitives (to limit the example to that given) as he would in a traditional concordance. His task is facilitated, however, as the editor says, by the prominent position of the keyword.[4] We may therefore say that the reasons given by the editors in question boil down to one, namely that if their authors used certain formulaic expressions or word combinations frequently, they would stand out immediately to the user, who merely has to glance down the page.

No one will dispute the interest such repeated expressions may have for researchers, or their possible importance to studies of linguistic peculiarities in their authors. Rather, the questions I would ask are: (1) how important are they with respect to the author in question? (2) how often do such repeated expressions occur? It would seem that the would-be editor must be able to answer 'very' to both questions, if he is to justify his adoption of this innovation. The editor must be the judge of the first question, in the first place; other specialists will also need to judge after the event. The second question, however, is a matter not of exclusive relevance to this author or that, but rather has to do with the principles of scholarship.

If a researcher sets out to find an answer to a question, and the results are only important if the answer is positive (that being his justification), he does not publish the results if they are slim or if the answer is negative. In some KWIC concordances - for example the *Baudelaire* and the *du Bellay* - the number or proportion of formulaic expressions or repeated word patterns is very slender, and it is difficult to see them constituting alone a justification of that method. (If the editor judges them to be important, even though they occur very little, their listing in a separate index might be a solution.) To repeat: the publication of results which do not support the stated justification of a method seems a questionable procedure. However, we must make distinctions. A work which comprises different parts or books, or which was written at different periods or stages of the author's life (for example Livy's *History*: see the example given by Packard (1968) in his Introduction), will in all likelihood be a more promising candidate for KWIC than a short work or one written in a brief time period.

The KWIC format is one answer to the particular problems of producing prose concordances. The earliest prose concordance was produced after the pattern of the traditional poetry concordance: the typographical line was taken as the context of each keyword it contained. It was soon seen, however, that this did not provide a sufficient citation to show the words in their context. The KWIC format assuredly provides a generous citation in the concordances where it is limited only by the size of the page. In two examples cited, the Packard *Livy* and the Hanon *du Bellay*, the citations are of 136 and about 120 characters, respectively. The trouble is, however, that these citations give false contexts. Since the citations are determined in an arbitrary manner, there is no guarantee that any given word will be seen in its true context, because the syntactical integrity of the writer's

language is not respected; but, on the other hand, one may be sure that there will be a considerable volume of superfluous matter. These defects will always be endemic to KWIC and constitute the grounds of the most serious criticism against them.»

It will not have passed unnoticed that in KWIC concordances, if they are to afford the maximum utility for which they have been designed, the frequency of occurrence of any word in an author's vocabulary is of prime importance. Any word which occurs only once or twice, or even ten or twelve times, will be of no significance for the study of formulaic expressions and word patterns. Such words, of course, comprise the vast bulk of any author's vocabulary. Identical phrases, formulaic expressions, and repeated word patterns, in order to be of any potential significance or to afford any material for separate worthwhile study, must be introduced by the commonest, most frequent words. It is their inclusion above all - not to mention the considerable amount of excess material reproduced by the KWIC context - which often results in volumes of inordinate size. The would-be editor must examine the question seriously, before embarking on his project, whether the obvious defects are going to be compensated for by results which may be of only marginal value.[5]

Given, on the one hand, the difficulty of reconciling the limitations of the computer with the complexities of an unedited prose text, and on the other the needs of the concordance user, the deviser and later users of KWIC have seen it as a means of maximizing the use of the computer and reducing to a minimum the tasks of the editor. This method has two significant virtues in its favour which every literary and linguistic scholar who depends on computers would like to attain: simplicity, and minimization of the human element - for the good reasons that they promote speed and reduce error.

To sum up, the KWIC method offers the following advantages: (1) the detection of formulaic expressions and repeated word patterns; (2) simplicity; (3) speed; (4) maximization of computer use; (5) minimization of the editor's role. The sceptical reviewer may be excused, in some cases, for wondering about the real order of precedence in the editor's mind.

All - or most?

We have seen that the KWIC concordance necessarily requires that it be exhaustive: that every word be entered and that all occurrences of every word be accompanied by citations.[6] What are the arguments in favour of completeness in a concordance if it is of the traditional type? The researcher who has, to my knowledge, both published such a concordance and written most forcefully in its favour is Professor Marvin Spevak.[7] His arguments are: (1) it should be done because the computer can do it 'simply and quickly'; (2a) it is a mistake to assume that 'frequency is directly proportional to unimportance'; (2b) most concordances do a disservice to scholarship 'by favouring only certain literary interests, and by almost neglecting sizeable linguistic interests'.

Perhaps a valid case can be made for a complete concordance to Shakespeare. But can a valid case be made for every writer - even for every major writer (however 'major' may be defined) - or for every work which migh merit a concordance? Anyone proposing to make and publish a concordance, and intending that it should be complete, would do well to calculate the sheer additional bulk of printed material involved, and weigh it against the expected increase in value and usefulness it will have for researchers.

Let me illustrate. The examples I give below come from Rabelais (I have chosen the shortest of his five Books). The twenty most frequent words in that text, in alphabetical order, are these:

à	598	l'	499
au	220	la	735
d'	385	le	697
de	1,568	les	622
des	357	n'	220
du	213	ne	373
en	979	par	399
et	1,418	qu'	251
il	238	que	638
ils ilz }	192	vous	228

making a total of 10,820. The total number of word tokens in the text is 38,276; and the total number of different words is 8,336. An exhaustive concordance of this Book would therefore comprise 38,276 complete citations. But of these no fewer than 10,820 would be accounted for by the twenty commonest words. Consider the proportions: nearly one-third of the whole for twenty words out of 8,336.[8] Here we are only considering twenty words, whereas there are in this text, as with all texts in English and French, perhaps sixty words which might well fall under the editorial axe. If sacrificial cuts must be made somewhere, they must be made at the expense of the commonest words. No editor would omit any of the words which occurred only once in his text. But he might well omit a word which occurred 10,000 times, or 1,000 times. The word 'and' occurs in Shakespeare 26,285 times. They are all there, in Spevack's Concordance, each with its own citation. They are there because 'in many instances omission requires more effort in the way of programming, if not computing time, than completeness'; and because there may be a number of scholars wishing to study the functional words in Shakespeare. (I suspect also - sympathetically - that completeness in itself affords great intellectual satisfaction.) But 'and' and a few other words constitute an immense volume of material which will never be studied by the majority user, the philologist, and the cost of which effectively puts its purchase out of his reach. And his editorial decision would be tantamount to saying that frequency is proportional to unimportance. But 'unimportant' does not mean totally useless or insignificant: it means relatively unimportant.

Why must omissions be made? Because the economic factor dictates it. Few will be convinced by Spevack's argument when he writes: 'Who would buy a thousand pages when he can buy six hundred? But this is not the place to argue the economics of the subject and the attendant demeaning of both project and public. What must be mentioned is our attempt to justify economic pressures with reasons which ... impair our goals'.[9] If it were only a question of a thousand pages against six hundred there would be no argument. It is rather a question of thousands of pages against 1,000-2,000 pages - of several volumes against a single volume.[10] For the aim should be to produce a one-volume concordance at a price which makes it accessible to the individual purchaser, to the scholars who are going to use it the most. If future editors and their publishers decide to publish complete concordances, they will find that even university libraries will discontinue buying them because of their high cost and rising costs.

Is there a way out of this dilemma? Do we want to find a way out? For two groups, of course, there is no dilemma: either a complete concordance, or incomplete and forget about the omitted words. For a third group, those who agree with Spevack that the functional words are of real interest to an important body of scholars, there is a dilemma, and it consists in finding the best way(s) to meet their needs.

One is to make the tapes available to researchers who have access to a computer. Another is to print the references only in an appendix to the main concordance. A third is to issue sub-concordances on microfilm or microfiche. A fourth - my

preference because of the greater convenience they would afford to the individual scholar - is a series of sub-concordances published as fascicules in the cheapest possible form. Since there is no compelling reason why each fascicule should contain a standard number of pages, their contents could be decided according to various possible criteria. Perhaps one reasonable way would be on the basis of grammatical categories - although such an arrangement would not be strictly accurate, since the English 'that' and the French 'que' (to cite only one example) would cut across more than one category. Or perhaps alphabetical - for ease of location.

Conclusion

In the foregoing analysis of two problems which must be resolved by every editor of a computer-generated concordance during its planning, and before he sets to work - namely, the problems of the citation and of completeness - I have tried to set forth some of the reasons which led me to conclude that the KWIC citation and the exhaustive concordance are unsatisfactory. They are both serious attemps to resolve the difficulties involved. And they both stem, I believe, from the desire of their various editors to meet the needs of two groups of scholars whose chief interests are different: the philologist and the linguist. The concordance (if we may distort history a little) was designed by and for the philologist; and the concordances so conceived and made have been, if not perfect, at least highly satisfactory. The KWIC and exhaustive concordances have been produced by scholars who, if not exclusively or chiefly linguists, have attached considerable importance to their needs. As Spevack has written: 'It may be interesting to note that the correspondence I have had regarding my concordance has been overwhelmingly from linguists, who want to know, first, if my complete concordance is really complete'.[11] And Kato: 'The present concordance has been edited with the conviction that in a concordance to a work of earlier date, exhaustiveness counts much, especially from a linguistic point of view'.[12] Hanon, in her Preface, says much the same thing. Perhaps the concordance is not the research instrument that really best meets the needs of linguists. Suffice it to say, however, that the traditional concordance, tested and proved by use during the past two hundred years, and improved by the refinements and analytical devices which the editor and computer can together provide, will continue to serve the philologist well in the future.

Notes

1. For example: *A Concordance to Livy*, edited by David W. Packard (Harvard University Press: Cambridge, 1968); *La Deffence et Illustration de la langue francoyse* by Joachim du Bellay. Concordance établie par Suzanne Hanon (Odense University Press, 1974). However, the context may be of variable length, as in: *A Concordance to Baudelaire's 'Les Fleurs du Mal'*, edited by Robert T. Cargo (University of North Carolina Press: Chapel Hill, 1965); *A Concordance to the Works of Sir Thomas Malory*, edited by Tomomi Kato (Tokyo University Press, 1974). I use the word 'citation' (as in the sentence to which this note is appended) to mean any passage of a text delimited so as to include a given headword or keyword; and the word 'context' to mean the natural syntactical unit of a given headword or keyword. This distinction is not made by all writers, some tending to use 'context' for both definitions.

2. This is so in Packard's *Livy*, Cargo's *Baudelaire*, and Hanon's *du Bellay*, but not in the *Malory*.

3. Cargo gives no explanation of his choice, except to say (Introduction, p. xiii) that this arrangement of materials makes for 'the optimum utility'. He adds: 'To my knowledge this arrangement of materials has never been used before in a published concordance ... the presentation of material, then, makes the Baudelaire concordance unique'.

4. One wonders, however, why the user would find it helpful to be able to locate the keyword rapidly in its surrounding milieu, since he must in any case read the whole citation. It is precisely because the KWIC citations are not limited to the briefest possible meaningful contexts, but reproduce much unnecessary text. The KWIC citation wastes the user's time, since he must read backwards and forwards from the keyword to find its syntactical context - which often is not there at all. The fact is, of course, that the KWIC concordance sets out to fulfil two irreconcilable aims, and sacrifices one to the other.

5. The most profitable use of a KWIC concordance necessitates knowing the words of greatest frequency in the author's texts. None of the volumes examined contains ranking lists or indexes of frequency.

6. Once again it should be stressed that different, and cheaper solutions to this feature, if it be deemed necessary, are either a limited-context concordance appended to the concordance proper, or simply an index of the formulaic expressions, etc., with references only or with limited contexts and references. If either of these courses were adopted, the chief justification for the KWIC layout in the main concordance would then disappear, of course.

7. See Marvin Spevack, 'Concordances: Old and New', in *Computer Studies in the Humanities and Verbal Behaviour*, IV, No. 1 (June, 1973), 17-19. The quotations that follow are, respectively, from pp. 17, 18 and 19.

8. The precise ratio is 28.2%. The equivalent figures in Shakespeare (taken from Spevack) are the following, in descending order of frequency:

the	27,457	is	9,603
and	26,285	not	8,992
I	21,206	me	8,218
to	19,938	for	8,100
of	17,079	it	8,052
a	14,675	with	7,834
you	14,326	his	7,210
my	13,075	this	7,074
that	11,725	your	7,003
in	11,511	he	6,683

giving a total of 256,046. Shakespeare's vocabulary comprises 29,066 different words; and his complete corpus numbers 884,647 words. The proportions are: the twenty commonest words, of a total of over 29,000, account for 28.9% of the volume of the concordance.

9. Art. cit., p. 18.

10. I should like to insist on the desirability of publishing a concordance in a single volume wherever it is feasible, for the important reason of the user's practical convenience. To have on one's desk, and to try to use, several volumes at once is more than a severe inconvenience: it is a physical impossibility. A multi-volume concordance will therefore inevitably restrict, and restrict seriously, the usefulness of the concordance.

11. Art. cit., p. 18

12. Loc. cit.